Lang Resources & Evaluation (2016) 50:585–601
DOI 10.1007/s10579-015-9302-y

ORIGINAL PAPER

AGH corpus of Polish speech

Piotr Żelasko[1] · Bartosz Ziółko[1,2] · Tomasz Jadczyk[1,2] · Dawid Skurzok[1,2]

Published online: 6 May 2015

Abstract A corpus of Polish speech, which has been collected for the purpose of automatic speech recognition (ASR) and text-to-speech (TTS) systems applications, is presented. The corpus consists of several groups of recordings: read sentences, spoken commands, a phonetically balanced TTS training corpus, telephonic speech and others. In summary duration of recordings is above 25 h. Number of unique speakers amounts to 166. The majority of them being in an age group of 20–35 and one third of them being female. Analysis of unique word occurrence frequency in relation to larger text resources has been concluded. From them, most commonly appearing words have been found and presented. The corpus was used as training data for the ASR system. Results of cross-validation training and testing the SARMATA ASR system using our corpus have shown that phrase recognition rate is 91.9 %. The corpus was additionally evaluated in comparative test against the CORPORA corpus, which had shown major increase in phrase recognition rate in favour of our corpus.

✉ Piotr Żelasko
pzelasko@agh.edu.pl;
http://dsp.agh.edu.pl

Bartosz Ziółko
bziolko@agh.edu.pl;
http://techmo.pl; http://dsp.agh.edu.pl

Tomasz Jadczyk
jadczyk@agh.edu.pl;
http://techmo.pl; http://dsp.agh.edu.pl

Dawid Skurzok
skurzok@agh.edu.pl;
http://techmo.pl; http://dsp.agh.edu.pl

1 AGH University of Science and Technology, Kraków, Poland

2 Techmo sp. z o.o., Kraków, Poland

Keywords Automatic speech recognition · Slavic languages · Polish · Speech corpus · Text to speech

1 Introduction

Research on automatic speech recognition (ASR) started about half a century ago (Halle and Stevens 1962; Denes and Mathews 1960; Denes 1960). Most of the progress in the field was done for English language. It has resulted in many successful designs. Still, ASR systems are below the level of human speech recognition capability, even for English. In case of less popular languages, like Polish (with around 60 million speakers), the situation is much more complicated. There are some commercial call centre applications, for example ones developed by PrimeSpeech, but they are strongly limited to their domain areas. There is no large vocabulary ASR (LVR) software for Polish, nonetheless several attempts have been made (Demenko et al. 2008; Pawlaczyk and Bosky 2009; Pułka and Kłosowski 2008; Marasek et al. 2009; Ziółko et al. 2011). Polish speech contains very high frequency phones (fricatives and plosives) and the language is highly inflected and non-positional. One of the main issues blocking development of such systems was lack of corpora of appropriate quality and size. The recent successes of the industry standard—the Polish ASR of Google, which is used for example in voice searching in Android applications, shows that this factor is highly determinative.

In order to develop ASR systems, a database of tens, even hundreds of hours of recorded speech is required for both training and testing purposes. For English language, such resources are well known and widely available—examples are the TIMIT corpus (Garofolo et al. 1993) or the Switchboard corpus (Godfrey and Hollman 1993). However, as we pointed out, for Polish, similar databases are scarce—there are few corpora of recorded speech, and very few available to be licensed. Recently, with new resources such as the Jurisdic corpus (Demenko et al. 2008) being available, Polish gradually ceases to be under-resourced. Yet, it still lacks other corpora which have hundreds of hours of properly annotated recordings. It also does not have a high quality phone-level manually annotated corpus which is at least few hours long, like the TIMIT corpus. Therefore, for the purpose of this paper, we will still consider Polish to be under-resourced in terms of speech corpora, but will not make such claims about Polish language resources in general. During our work on various projects, we managed to collect some sets of recordings which are not too comprehensive by themselves, but put together in one corpus should be large enough to satisfy the needs of ASR system training, therefore, hopefully, adding another resource to the pool.

As a first step, we would like to discuss the problems of ASR systems development for under-resourced languages. Then, shortly present the major available corpora for the Polish language (we have been using some of them in our previous and current works). We will also show an example of annotation and describe how the corpus is annotated. In next step we will describe each of the major elements of our corpus and introduce its statistics, such as the speakers' gender and age or total number of words (also unique words) along with a list of the most frequent ones. We will also show how the corpus is useful in performing ASR system training and tests along with the results of its application in the SARMATA ASR system (Ziółko et al. 2011).

1.1 ASR for under-resourced languages

Only a handful of the world's languages, like English, benefit from resources such as wide selection of hundreds of hours long speech corpora or representative text corpora. Those are useful in a number of speech-associated applications, such as speech recognition, speaker identification and verification or interactive dialogue system development (Scannell 2007). Therefore, most of the world's languages should be qualified as under-resourced, and Polish is one of them, at least in terms of high quality speech corpora availability.

The main problem of creating language resources is the high cost of their production. The work associated with process of language resources creation is mostly manual, which is both costly and not very effective—e.g. manual transcription can take several times as long as the recordings' duration and manual annotation on the phone level takes even tens of times longer. On the other hand, to satisfy the needs of statistical models employed in modern speech-related technologies, a large amount of training data is required. It makes corpus creation a very difficult task for researchers, who wish to provide such data.

In case of text corpora, collecting text data from the Internet mitigates some of these problems. The Internet is a source that seems to be almost unlimited in size, and it also offers wide variety of resource types: social media such as Facebook or Twitter offer short, colloquial texts, while blogs and news portals might offer more formal and longer texts. Also, a lot of literature has been already converted to digital form and are freely available on the Internet. It has been shown that corpora built on Internet resources bring promising results (Scannell 2007; Kilgarriff and Grefenstette 2001). Examples of works which utilized the Internet in building a language resource or using it for some purpose are an attempt to build clean bilingual corpus (Resnik 1999), an n-gram model (Ziółko and Skurzok 2011) or even a model for irony detection in short texts (Reyes et al. 2013).

This is, however, not the case in construction of speech corpora. Although lots of recorded speech can be found on the Internet, it is almost never transcribed, which rules out the ability of supervised training or evaluation of a system under development. Another problem with freely available speech recordings is their quality—most people do not own a high quality recording device, and so the resulting recordings tend to feature any combination of the following qualities: distortion, narrow spectrum, high levels of noise, both environmental and originating from the recording device itself, dynamics compression, effects of encoding the signal using lossy codec such as MP3 and others. While some of these features may be useful to have in a corpus dedicated to some particular application (e.g. low-quality telephonic recordings for a telephonic ASR system testing), they are generally obstacles in development of speech-related applications. There are also legal issues that need to be covered when using a recording of somebody's speech, which vary among different countries. Asking the speaker for permission to use his voice might be a severe obstacle in automatic retrieval of speech samples from the Internet.

Therefore, the approaches which are left for development of speech databases are:

- creation of totally new set of annotated recordings, which meet certain requirements established by the corpus designer (e.g. phonetic variety, high number of unique speakers, continuous speech, etc.), and
- adaptation of available collections of recordings, such as recordings from call-centres or public speeches and lectures, which most often involves creating a manual annotation of time boundaries of recorded utterances.

An example of the first kind of corpus for Polish is CORPORA (Grocholewski 1997) and an example of the second one is LUNA (Marciniak 2010)—both are described in the next section.

It needs to be mentioned that some researchers found a way to deal with lack of resources during construction of phonetic models for purpose of ASR by means of bootstrapping (Schultz and Waibel 1997; Le and Besacier 2009). The idea is to transfer models created for phonemes of one language to another, with some kind of transformation or retraining involved. The results of this technique seem promising, and comes with a great advantage which does not require as much resources as traditional training techniques. It still has to be considered that an annotated corpus is needed for the evaluation of the resulting ASR system.

1.2 Known corpora of Polish speech

The most well-known speech corpus for Polish is CORPORA, prepared in 1993 by Grocholewski (1997). It contains 365 utterances (numbers, names and 114 sentences) spoken by 45 speakers. The sentences are incoherent semantically, due to provide maximal phonetic diversity. An example of such a sentence is "on myje wróble w zoo", which in English means roughly "he is washing sparrows in a zoo". Recordings of two speakers (a woman and a man) were manually annotated to phonemes and then, with help of dynamic programming methods, the rest of the speakers were annotated automatically.

A relatively large corpus of speech related to legal matters is named Jurisdic (Demenko et al. 2008). It contains recordings of about 1000 different speakers from different parts of Poland. Half of the recordings come from Court, Police, Prosecution and the other half from universities and offices. Each speaker recorded about half an hour of both spontaneous and read speech.

LUNA is a corpus of spontaneous telephonic speech (Marciniak 2010). It contains about 11 h of dialogues, where one person—the caller—asks for information concerning public transport in Warsaw, and the other person—the agent—gives that information. Originally, the corpus had only morphosyntactic and semantic annotation. A time annotation to words was made separately for training of SARMATA ASR system (Ziółko et al. 2011).

NKJP stands for "Narodowy Korpus Języka Polskiego"—in English, "National Corpus of Polish" (Przepiórkowski et al. 2012). It is a large resource of Polish texts, which consists of literature, journalism, letters, Internet texts and others. It is also a resource of recorded conversations and media speeches, which unfortunately, are not provided with time aligned annotation.

Recordings of Polish speech are also available as parts of larger, multilingual corpora. GlobalPhone is a corpus consisting of 20 languages, where for each language 100 sentences were read by 100 speakers (Schultz 2002). Chosen texts focus mostly on national and international politics as well as economics news.

Another database is the SpeechDat-E, which contains recorded speech from fixed telephone networks for five eastern European languages, including Polish. There are recordings of 1000 speakers, each reading some sentences and isolated words.

There is also a corpus consisting of recordings from the Europarliament, which is claimed by its authors to have several hundreds of hours of Polish politicians and interpreters speech. Unfortunately, this data has only approximate transcriptions or none at all (Lööf et al. 2009).

EASR is a corpus of elderly speech for 4 languages, Polish amongst them (Hämäläinen et al. 2014). Authors claim to have 205 h of read speech of 781 Polish persons.

The paper does not cover all speech recordings made by the DSP AGH group. It should be mentioned that our others, more specific speech corpora were already described: audio–visual speech corpus of Polish and a corpus of Polish emotional speech. Hopefully, a corpus designed for Polish speaker identification and verification will also soon be described and published. For more information on these corpora and their availability, do not hesistate to contact us.

1.3 Master label file format

The AGH corpus is annotated with master label file (MLF) format. The kind of annotation we use carries information about beginning and ending times of either a word or a phrase. It may also be used to mark range of time in which an agent (e.g. human transcriber, ASR system) perceives the basic unit of speech—a phone. Files containing the annotation have .mlf extension and must contain a proper header, which consists of #!*MLF*!# symbol in the first line. Multiple annotations may be contained in one MLF, established that each annotation begins with a path to annotated wave file and ends with a dot. Basic time unit in MLF is 100 ns, that gives precision which is much better than typically needed in any ASR system. The basic time unit in our annotations is 1 ms. The MLF format was originally designed for the HTK software (Young et al. 2005). Below, an example of annotation in MLF format is shown.

```
#!MLF!#  # The header.
''waves/spk_1/1.wav''  # The path to annotated file.

# A comment.

0 12490000 heja # Annotated word of 1249ms duration.
12890000 24310000 co_u_ciebie # Annotated phrase.

24310000 25720000 o
25720000 27260000 l
27260000 28190000 u # Annotated phonemes.

. # Dot signs the end of 1.wav file annotation.

''waves/spk_1/2.wav'' # Path to the next file.
#... Rest of the annotation.
```

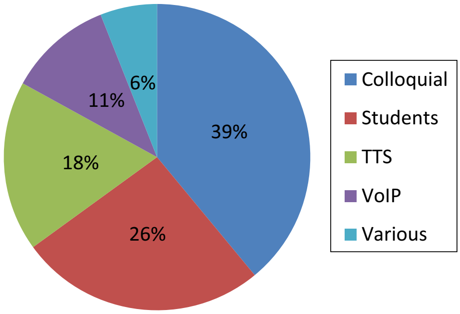
2 Contents of the corpus

We present a description of every major part of AGH corpus. Issues discussed are: how and why the recordings set was created, what kind of speech do the recordings contain, recording equipment and gender ratio of the speakers as well as a more general description.

The recorded speech is contained in a single channel (mono) WAVE files with sampling frequency of 16 kHz and 16 bit precision. Annotations were automatically checked for orthographic correctness using OpenSJP, an open source distribution of Polish dictionary (OpenSJP 2014) and manually corrected. Some parts (e.g. students' recordings) were also checked manually. The post-processing of the corpus included creation of a list of words which are foreign or phonetically ambiguous. Preparation of a transformation dictionary for them and transforming them in such a way, so that ORTFON [our software performing automatic phonetic transcriptions based on an algorithm created by Steffen-Batóg and Nowakowski (1992)] is able to deal with these problematic words. An example of such transformation: let us suppose that we have found an annotated phrase "earl grey", in which both words are foreign to Polish. In order to help ORTFON transcribe it properly, we change it to "erl grej", which is more phonetically compliant with Polish and allows the usage of the same transcription rules as for proper Polish words. The process is reversible because of creation of the transformation dictionary containing rules used for transformation, which also makes it possible to apply them to new data in the corpus as well. It also means that in this case the ASR system will output "erl grej" as the recognized phrase, although we can postprocess the results and convert it automatically to "earl grey".

The prevalent dialect in the corpus is the Lesser Polish dialect, because the majority of speakers come from the south-eastern regions of Poland. However, other dialects are also present. We did not mark information about the descent of each speaker because it is not needed for our current research. Still, we feel that this fact requires some explanation. Dialects in Polish are weaker than e.g. in English and generally avoided in formal situations. This fact is resembled in census collected by

Fig. 1 Percentage of each of subcorpus contribution to our corpus in terms of recordings length



Central Statistical Office of Poland in 2011, which indicates, that only 2 % of Poles use other languages or dialects at home (GUS 2011). Furthermore, recent research on tolerance of dialects in Poland suggests that although Poles declare to be tolerant of regional dialects, they prefer not to hear them and expect others not to use them in formal situations (Hansen 2014). Another example of dialect marginalization is that in "National Corpus of Polish" (Przepiórkowski et al. 2012) there is no mention of dialect itself. Therefore we did not consider tagging of dialects such significant during preparation of our corpus.

Our choices of the methods of data collection as well as decisions on the statistic profile of the corpus were mainly dictated by the need of large number of speakers and large amount of recordings (Fig. 1). We focused primarily on building large and well annotated training corpus rather than on representing complete set of various dialects, ages or topics. We designed our corpus to be dedicated to ASR training and tests, and therefore provided all required metadata only for those tasks.

2.1 Colloquial speech recordings

This set of recordings has been created by 10 speakers (5 males and 5 females), each one reading out about 1000 phrases in the near field of a microphone inside a small room with no audible reverberance, isolated from the outside noise. For every speaker, summary recordings length is about 1 h, resulting in total length of 10 h and 4 min. Each phrase is saved in a separate WAVE file and is annotated in an MLF file as a whole (beginning and ending of each phrase is marked, without distinguishing single words). Recorded utterances are short sentences or fragments of longer sentences, which are picked from the Internet. They are derived from everyday language and ensured to be orthographically correct. These recordings were prepared for us by an external company.

2.2 Students' recordings

One of the classes we teach, concerns the speech-dedicated technologies. In order to pass the course, the students are expected to build a simple purpose ASR system using HTK (Young et al. 2005) for their own voice. Examples of such projects are systems handling pizza ordering, tickets booking or providing a voice interface for some application. Two major steps leading towards creation of such systems are grammar design and preparation of recordings. The grammars prepared are simple and designed for recognition of one sentence, which contains all the relevant information (such as quantity, size and topping of ordered pizzas). For each project, about 3 min of recordings are prepared. Recorded utterances are sentences compliant with grammar obtained in the earlier step or enumerations of words from the dictionary. Later on, the recordings are annotated to words by students, and then converted to phoneme level annotation by ASR system SARMATA (Ziółko et al. 2011).

At the moment we have 125 students recorded and we expect this number to grow around 60 persons per each year. The distribution of gender is 86 males and 39 females, giving ratio of roughly 2:1. Also, the vast majority of speakers are in the age group of 20–25. Duration of this subcorpus is 6 h and 33 min. Equipment used to prepare these recordings (and so their quality) is various: some students use cheap PC microphones and cell phones, but some have professional recording equipment at their disposal, such as dictaphones or high-end microphones with suitable audio interfaces. This allows for testing how dependent the ASR is of recording devices.

2.3 TTS training corpus

During our efforts to develop a text-to-speech (TTS) synthesizer, a large set of recordings was prepared in order to be used by the system. It consists of 2132 sentences uttered by a young woman, who is a trained speaker. The text comes from the 1 million words subcorpus of NKJP corpus (see Sect. 1.2; Przepiórkowski et al. 2012) and was designed to be both phonetically rich and balanced. It means that the distribution of IPA phonemes and diphones occurrence frequency is as close to Polish language as possible. Accomplishing this condition is important for training and testing of both ASR systems and TTS synthesis (Abushariah et al. 2012). Total length of recordings amounts to 4 h and 30 min. The quality of the recordings is proper, as they were made in an anechoic chamber with high-end recording devices (Felis et al. 2012). Original recordings were sampled at 44.1 kHz to meet the expectations of speech synthesis, but for the purpose of ASR training and tests, we also prepared a downsampled 16 kHz version with the use of SoX software.

2.4 VoIP recordings

During development of interactive voice response (IVR) system dedicated to testing SARMATA ASRs (Ziółko et al. 2011) performance, we recorded various utterances which are keywords in voice menu navigation and also combinations of numbers. We put special attention to covering every possible number transition. Examples of utterances are “Internet”, “telefon” (phone), “usterka” (fault), “jeden sześć osiem dwa” (one six eight two). The recordings were obtained using a voice over IP (VoIP) application, which accepted calls from public switched telephone network (PSTN) phones and cell phones. They are stored in WAVE pulse code modulation (PCM) format, but they were previously encoded and decoded during transmission. Total number of speakers is 27, with 17 males and 10 females. Total length of recordings amounts to 2 h and 52 min. The majority of speakers are 20–35 years old. Recordings are annotated to words.

2.5 Various recordings

In the corpus, there are also other sets of recordings of lesser quantity. Some of them are recorded commands used for SARMATA ASR system (Ziółko et al. 2011) testing in specific scenarios, such as SAWA, a project of voice interface to a documentation database program used by institutions of justice, a virtual mouse project or a virtual advisor project. Other recordings are public lectures given by members of our group about speech technologies or presentations from our seminars. This subcorpus has 1 h and 39 min duration and is annotated to words.

We also have 15 min of a read text (3 speakers, each reading for about 5 min) manually transcribed to phones. For this task, we used a phonetic alphabet which is a simplification of speech assessment methods phonetic alphabet (SAMPA) for Polish (Ziółko et al. 2007). Speakers are two men and one woman in age 20–35 years old. All recordings mentioned in this subsection were done using wireless recording devices such as AKG WMS 40. The purpose of having some manual phone-level annotation is that it can be used to provide a starting point for training of phone models in the ASR system as well as research on phonemes lengths, varieties depending on accent and location in a sentence etc.

3 Corpus statistics

In total, AGH corpus has 25 h and 38 min of recordings, as presented in Table 1. It also has 166 speakers, and one third of them are female. Both males and females contribute equally to the corpus in terms of total recordings duration for each gender—although there are more men recorded, the TTS recordings feature 4.5 h of female voice. Majority of speakers are 20–35 years old.

The corpus contains 117,450 words. 13,784 words are unique, and about half of them appear only once. Also, about 14 % of these unique words appear 10 times or more. Fig. 2 illustrates how many words with specific numbers of occurrences can be found in the corpus, however, for sake of clarity, we removed the single occurrence words and trimmed the histogram to 250 word occurrences. It resulted in 39 most frequently appearing words (0.3 % of vocabulary in the corpus) being omitted. We also checked how many unique phrases may be found in the corpus. Because some parts of the corpus are annotated to words, we created an additional annotation to phrases, in which we joined the words which are separated by time interval not >100 ms. The corpus has 9440 such unique phrases (some of which are separate words).

In Table 2, we present the most frequent words in our corpus, along with their occurrence frequency, calculated as a percentage of the occurrence count of a given word with respect to summary word count in the corpus. We observed that many of the most frequently appearing words are prepositions, conjunctions and pronouns, which suggests that recorded sentences have natural syntax, even though it is often a simple one. These results are similar to ones observed in earlier works, where

Table 1 Contribution of each subcorpus in terms of recordings length

Corpus	Duration
Colloquial speech (2.1)	10 h 6 min
Students (2.2)	6 h 33 min
TTS (2.3)	4 h 30 min
VoIP (2.4)	2 h 52 min
Various (2.5)	1 h 39 min
Total	25 h 38 min

Numbers following the subcorpus' name indicate in which section its description can be found

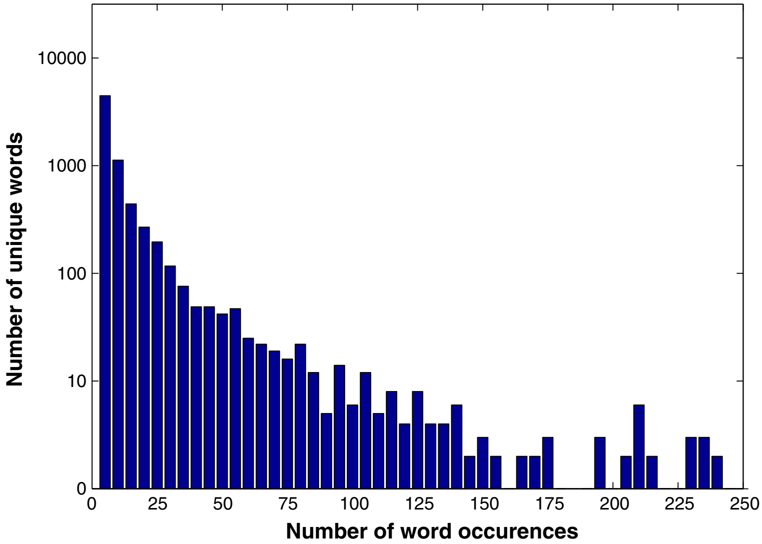


Fig. 2 Distribtuion of occurrence frequency of unique words in the corpus, presented using a logarithmic scale. Single words and words with count higher than 250 were omitted

occurrences of n-grams for Polish were investigated based on a 9 GB text corpus (above 1 billion words) consisting of newspapers, Wikipedia articles and literature (Ziółko and Skurzok 2011). In both corpora, these parts of speech are appearing at the top of the lists and have similar occurrence frequency. In the recorded corpus, some adjectives associated with giving commands to the ASR system, like "poproszę" (Eng. *I'd like*) or "kup" (Eng. *buy*) are much more frequent than in natural language. It is caused by their extensive usage in students' projects (see Sect. 2.2), which often offer similar functionality—e.g. both system handling a pizza order and system handling a coffee order will anticipate that user might begin the sentence with "I'd like".

Table 2 List of 60 most frequent words along with their translation to English and their occurrence frequency in our corpus and in a 1 billion words corpus from earlier works (Ziółko and Skurzok 2011)

Word	Translation	Occurrence in AGH corpus (%)	Occurrence in (Ziółko and Skurzok 2011; %)
na	on	2.55	1.67
w	in	2.38	2.26
z	with	1.89	1.51
się	*a*	1.54	2.39
i	and	1.48	2.34
do	to	1.30	1.10
nie	no	1.18	1.82
jest	is	0.94	0.43
o	at	0.90	0.54
to	this/it	0.87	0.98
poproszę	I'd like	0.83	0.00
jak	how	0.44	0.51
czy	*b*	0.41	0.24
co	what	0.37	0.40
włącz	turn on	0.37	0.00
bilet	ticket	0.36	0.00
dwa	two	0.35	0.05
że	that	0.34	0.95
po	after	0.33	0.44
a	and	0.33	0.67
pięć	five	0.32	0.02
proszę	please	0.31	0.04
mnie	me	0.30	0.22
za	*c*	0.29	0.35
dzisiaj	today	0.29	0.01
trzy	three	0.28	0.03
numer	number	0.27	0.01
są	they are	0.26	0.12
od	since	0.24	0.32
już	now	0.24	0.27
być	to be	0.23	0.11
później	later	0.23	0.04
jednak	however	0.23	0.14
cztery	four	0.23	0.02
może	maybe	0.22	0.19
sześć	six	0.22	0.01
ale	but	0.22	0.42
gdzie	where	0.22	0.08
mogę	can I	0.22	0.04
kup	buy	0.21	0.00

Table 2 continued

Word	Translation	Occurrence in AGH corpus (%)	Occurrence in (Ziółko and Skurzok 2011; %)
siedem	seven	0.21	0.01
osiem	eight	0.21	0.00
ma	has	0.21	0.14
u	at	0.20	0.08
połóż	put	0.20	0.00
wyłącz	turn off	0.20	0.00
przez	through	0.20	0.27
jeden	one	0.19	0.05
zero	zero	0.19	0.00

Numbers denoted as 0.00 are <0.005

a Reflexive pronoun;

b Particle used to create "yes or no" questions;

c Preposition with meanings: behind, for, after, in, for

In order to check if there is a correlation between occurrence of words in the recorded corpus and the text corpus, we calculated the Pearson's correlation coefficient r,

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}, \tag{1}$$

where x is a descendingly sorted vector of occurrence frequencies for each word in the recorded corpus, y is a vector of occurrence frequencies in the text corpus calculated for words from x without changing their order, n is the number of words in the recorded corpus and $\bar{x}$ and $\bar{y}$ are means for their respective vectors. Correlation of both data sets is high with $r = 0.88$, which indicates that although the recorded corpus consists of much less words than the text corpus, their distribution appears to be similar—that is, a word likely to be frequent in one corpus is probable to be frequent in the other one.

4 Corpus evaluation in ASR

We chose two methods of corpus evaluation with help of the SARMATA ASR system (Ziółko et al. 2011). The first one involves a cross-validation procedure using only our corpus. The second one involves ASR tests on GlobalPhone with system being trained exclusively on CORPORA, exclusively on our corpus and then on both of them joined. We begin with description of the training procedure, and then we describe both evaluation methods in detail.

4.1 Training of acoustic models

In order to perform parametrisation of the input signal, the system employs 20 ms time window segmentation with 10 ms overlap and computes mel frequency cepstral coefficients (MFCC) and frame energy along with their deltas and double deltas. Hidden Markov models (HMM) is being used as a classifier—each acoustic model is 3-state and uses a 20-Gaussian mixture. A set of 37 phones from SAMPA for Polish was used in the procedure. Phonetic transcriptions were automatically prepared by the ORTFON program.

Acoustic models are trained using SARMATA ASR. They are initialized by assigning in each model the same value of means and variances to the components of Gaussian mixture. To perform training, we use an expectation-maximization (EM) technique. In the first iteration, phonetic alignment is achieved by uniform segmentation of acoustic data (flat start). Results are used to fit the Gaussian mixtures in acoustic models. Next iterations use the previously trained models and the Viterbi algorithm to classify data. The training stops once convergence is achieved (i.e. no more improvement is gained by the next iteration).

4.2 Cross-validation

In the first evaluation scenario, we prepared 5 cross-validation sets of training data (80 % of corpus, about 20 h) and test data (remaining 20 % of corpus, about 5 h), so that in each set, the test data was unique. In the next stage, using each set, we trained the system and then conducted recognition tests. Because some parts of the corpus are annotated to words and others are annotated to phrases, we concatenated the words in every phrase, so that it became one longer word. In Table 3 we present the percentage of correct phrase recognitions. In this context, by phrases we mean both single words and groups of words concatenated to a single word. The mean phrase recognition rate is 91.9 % and its standard deviation is 1 %.

Results of tests in five cross-validation data sets are similar. Since each training and test set contains the same percentage of each of the subcorpus, the variation in the results cannot be explained by differences in recording quality, gender or dialect. The reason might be that some test sets (such as set 2) contain phrases which are less phonetically complex, and thus easier to recognize for the system, than other test sets (such as set 1).

Table 3 Recognition results for each data set in cross-validation testing procedure

Set	Phrase recognition rate (%)
1	91.0
2	93.4
3	91.1
4	91.4
5	92.4

The results shown are percentages of correctly recognized phrases

4.3 Tests on other corpora

The second evaluation scenario is to check how well the system performs when trained and tested on recordings from different corpora. For this purpose, we trained three variants of the system: using CORPORA (Grocholewski 1997; 14,940 utterances, 4 h 54 min), using our corpus and using both corpora. In order to test the outcome, we took 16 randomly chosen speakers from the Polish part of GlobalPhone (Schultz 2002; 4583 utterances, 3 h 58 min) as a test corpus. The speakers chosen for the test had numbers: 2, 6, 7, 15, 16, 17, 21, 26, 48, 53, 58, 59, 78, 82, 84 and 87. All of their recordings are supplied with sentence-level time alignment.

Before we discuss the results of tests, we will describe some features of the GlobalPhone corpus which may affect the ASR performance. It should be noted that in the GlobalPhone recordings, different types of noise are present. There is some static noise, which is suspected to be generated by the recording device, as well as noisy events, such as hitting the microphone, blowing in the microphone (i.e. popping) and environmental noise. Moreover, the recorded speech is full of disfluencies, such as silent pauses, breath pauses, interjections (fillers), revisions, repetitions and others (these phenomena are described in detail e.g. in Rochester 1973; Fromkin 1984), which are not present in our model, and so force the ASR to align some phone in their place. Given recordings of this quality, the system is expected to yield worse results.

Training the system on CORPORA resulted in very poor recognition rate. We suspect several reasons behind this outcome (Table 4). Firstly, although CORPORA is a phonetically rich corpus, the phrases are designed to include rare phonetical contexts which are difficult to pronounce. This leads to non-natural and error-prone pronunciation, which is different from that in normal read speech and spontaneous speech. It also seems that 5 h of training data is not enough to allow solid training of acoustic models as described in Sect. 4.1. Finally, CORPORA recordings are of better quality and recorded under different conditions than those in GlobalPhone, so the trained models might be mismatched.

Using our corpus to train the ASR yielded recognition rate almost 30 % points better than when CORPORA was used. We attribute such improvement in performance mostly to the increase in amount of training data, which allowed the acoustic models to converge better. We suspect that the variety of recording environments and recording equipment present in our corpus also helped by making the acoustic models more general and appropriate in more use cases, and therefore

Table 4 Recognition results for tests concluded on GlobalPhone with system being trained separately on CORPORA, AGH corpus and their sum

Training data	Testing corpus	Phrase recognition rate (%)
CORPORA	GlobalPhone	54.9
AGH	GlobalPhone	83.7
CORPORA + AGH	GlobalPhone	84.2

The results shown are percentages of correctly recognized phrases better suited to recognize GlobalPhone speech than CORPORA trained-models were. The performance is, still, worse than when both training and testing the system with our corpus. We believe it can be explained by the quality of GlobalPhone recordings, as described earlier in this section.

The concatenation of our corpus and CORPORA for training purposes allowed the SARMATA system to further improve the recognition rate by about 0.5 % point. We believe that the addition of CORPORA helped in training acoustic models of these phones, which are rare in our corpus.

5 Conclusions

AGH corpus is one of the largest documented corpora of Polish (over 25 h and 166 speakers), featuring a variety of speech scenarios, including text reading, issuing commands, telephonic speech, phonetically balanced 4.5 h subcorpus recorded in an anechoic chamber and others. The resulting corpus is a collection of varied recordings, which originally served different purposes, but were gathered together to create one of the largest resources of this kind for Polish, targeted at ASR system development. The speech recorded in the corpus mentions, but does not focus on any specific area such as politics, economics or law. Most parts of the corpus are semantically coherent sentences built using simple, natural syntax. Foreign and phonetically ambiguous words were corrected to improve automatic phonetic transcription. The word occurrence frequency is correlated with that of a large text corpus, representative for Polish.

As has been shown, the corpus under description is in fact a concatenation of several smaller corpora, which we acquired during our work. It results in quite broad coverage of types of speech (e.g. isolated commands or fluent sentences) and recording conditions (i.e. cheap or high-end microphones and telephones, noisy or quiet rooms), but is not targeted at any specific application, such as telephonic dialogue systems or interfaces to computer programs or electronic devices. This lack of corpus specialization may be seen both as a strength and as a weakness of the corpus. The strength is that the corpus may be used as a starting point for a broad range of applications, and the weakness is that in each application additional data is needed in order to develop the ASR. For example, in order to train the ASR system for a telephonic dialogue system application, whole corpus may be used to train a general model and then, an application-specific, smaller corpus, which consists of a target vocabulary and telephonic speech only, could be used to retrain the system, resulting in a more specialized model. From another point of view, thanks to the fact that our data has different origins, the corpus achieves betters representativeness in terms of various speech contexts and recording environments, even if the same cannot be said in terms of e.g. age distribution of speakers or dialects of Polish. Also, our policy of accepting any type of recording helped in gathering larger number of speakers, which leads towards achieving better speaker independence of trained models.

Our corpus was designed primarily for ASR training and it has been shown to perform well in this task. It was used to train SARMATA ASR system, helping it achieve 91.9 % of correct phrase recognitions during cross-validation on our corpus alone. The corpus was additionally evaluated by separately training the acoustic models on CORPORA and the AGH corpus. In comparative tests on GlobalPhone, our corpus outperformed CORPORA with almost 30 % points difference in the recognition rate. Also, as shown in results of our tests, varied sources of training data combined with higher quantity of speakers indeed made the acoustic models more representative than data originating from CORPORA, where context, recording room and equipment are the same for each utterance. Therefore, our methodology of gathering as much data as we could and combining it into a single resource specialised in ASR training appears to be successful.

We expect the AGH corpus to grow over time and to introduce new speakers and new types of recording scenarios. The corpus is available on both academic and commercial licence.

Acknowledgments We would like to thank all whose voices are in the AGH corpus. We also thank Anna Grabska for revising English in this paper. This work was supported by LIDER/37/69/L-3/11/NCBR/2012 grant.

References

Abushariah, M., Ainon, R., Zainuddin, R., Elshafei, M., & Khalifa, O. (2012). Phonetically rich and balanced text and speech corpora for Arabic language. *Language Resources and Evaluation*, *46*(4), 601–634.

Demenko, G., Grocholewski, S., Klessa, K., Ogórkiewicz, J., Wagner, A., Lange, M., Śledziński, D., & Cylwik, N. (2008). JURISDIC: Polish speech database for taking dictation of legal texts. *Proceedings of the International Conference on Language Resources and Evaluation* (pp. 1280–1287).

Denes, P. (1960). *Automatic speech recognition: Experiments with a recogniser using linguistic statistics*. Technical report, DTIC document.

Denes, P., & Mathews, M. (1960). Spoken digit recognition using time-frequency pattern matching. *The Journal of Acoustical Society of America*, *32*(11), 1450–1455.

Felis, J., Flach, A., & Kamisiński, T. (2012). Testing of a device for positioning measuring microphones in anechoic and reverberation chambers. *Archives of Acoustics*, *37*, 245–250.

Fromkin, V. (1984). *Speech errors as linguistic evidence. Janua Linguarum. Series maior*. Berlin: De Gruyter.

Garofolo, J., Lamel, L., Fisher, W., Fiscus, J., Pallett, D., Dahlgren, N., et al. (1993). *TIMIT acoustic-phonetic continuous speech corpus LDC93S1*. Philadelphia: Linguistic Data Consortium.

Godfrey, J., & Hollman, E. (1993). *Switchboard-1 release 2 LDC97S62*. Philadelphia: Linguistic Data Consortium.

Grocholewski, S. (1997). CORPORA-speech database for Polish diphones. *Proceedings of Eurospeech*.

GUS. (2011). Ludność Stan i struktura demograficzno-spoeczna. Narodowy Spis Powszechny Ludności i Mieszkań 2011.

Halle, M., & Stevens, K. (1962). Speech recognition: A model and a program for research. *IRE Transactions on Information Theory*, *8*(2), 155–159.

Hämäläinen, A., Avelar, J., Rodrigues, S., Dias, M. S., Kolesiński, A., Fegyó, T., Németh, G., Csobánka, P., Lan, K., & Hewson, D. (2014). The EASR corpora of European Portuguese, French, Hungarian and Polish elderly speech. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Hansen, K. (2014). Stosunek Polaków do dialektów regionalnych. raport na podstawie Polskiego Sondażu Uprzedzeń 2013.

Kilgarriff, A., & Grefenstette, G. (2001). Web as corpus. In *Lancaster University* (pp. 342–344).

Le, V.-B., & Besacier, L. (2009). Automatic speech recognition for under-resourced languages: Application to Vietnamese language. *IEEE Transactions on Audio, Speech, and Language Processing*, *17*(8), 1471–1482.

Lööf, J., Gollan, C., & Ney, H. (2009). Cross-language bootstrapping for unsupervised acoustic model training: Rapid development of a Polish speech recognition system. *Proceedings of Interspeech, Brighton* (pp. 88–91).

Marasek, K., Brocki, Ł., Koržinek, D., Szklanny, K., & Gubrynowicz, R. (2009). User-centered design for a voice portal. *Aspects of Natural Language Processing, Lecture Notes in Computer Science*, *5070*, 273–293.

Marciniak, M. (Ed.). (2010). *Anotowany korpus dialogów telefonicznych*. Warsaw: Akademicka Oficyna Wydawnicza EXIT.

OpenSJP (2014). Open source online dictionary of the Polish language. http://sjp.pl. Accessed 10 Apr 2014.

Pawlaczyk, L., & Bosky, P. (2009). Skrybot: a system for automatic speech recognition of Polish language. *Advances in Soft Computing, Man-Machine Interactions, Springer*, *59*(2009), 381–387.

Przepiórkowski, A., Bańko, M., Górski, R., & Lewandowska-Tomaszczyk, B. (2012). *Narodowy Korpus Jezyka Polskiego*. Warsaw: Wydawnictwo Naukowe PWN.

Pułka, A., & Kłosowski, P. (2008). Polish semantic speech recognition expert system supporting electronic design system. *Proceedings of Conference on Human System Interactions (HSI), Krakow* (pp. 479–484).

Resnik, P. (1999). Mining the web for bilingual text. In *In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics* (pp. 527–534).

Reyes, A., Rosso, P., & Veale, T. (2013). A multidimensional approach for detecting irony in Twitter. *Language Resources and Evaluation*, *47*(1), 239–268.

Rochester, S. (1973). The significance of pauses in spontaneous speech. *Journal of Psycholinguistic Research*, *2*, 51–81.

Scannell, K. P. (2007). The crúbadán project: Corpus building for under-resourced languages. In *Building and Exploring Web Corpora: Proceedings of the 3rd Web as Corpus Workshop* (Vol. 4, pp. 5–15).

Schultz, T. (2002). Globalphone: A multilingual speech and text database developed at Karlsruhe University. In *Proceedings of the ICSLP* (pp. 345–348).

Schultz, T., & Waibel, A. (1997). Fast bootstrapping of LVCSR systems with multilingual phoneme sets. In *Proceedings of Eurospeech, Rhodes* (pp. 371–374).

Steffen-Batóg, M., & Nowakowski, P. (1992). An algorithm for phonetic transcription of orthographic texts in Polish. *Studia Phonetica Posnaniensia*, *3*, 135–183.

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Moore, G., et al. (2005). *HTK Book*. UK: Cambridge University Engineering Department.

Ziółko, B., & Skurzok, D. (2011). N-grams model for polish. In Ivo Ipsic (Ed.), *Speech and language technologies* (pp. 107–127). InTech.

Ziółko, B., Gałka, J., Manandhar, S., Wilson, R., & Ziółko, M. (2007). Triphone statistics for Polish language. *Proceedings of 3rd Language and Technology Conference, Poznań*.

Ziółko, M., Gałka, J., Ziółko, B., Jadczyk, T., Skurzok, D., & Mąsior, M. (2011). Automatic speech recognition system dedicated for Polish. *Proceedings of Interspeech, Florence*.